Athens University of Economics and Business

MSc in Data Science

Data Mining – Assignment 1

Deadline: 28/1/2020

Group assignment (groups of up to 2 people).

The assignment corresponds to 20% of the total grade of the course.

Discussions between groups are recommended but collaborating on the actual solutions is considered cheating and will be reported.

There will be no extension of the assignment deadline!

Professor: Y.Kotidis (kotidis@aueb.gr)

Assistant responsible for this assignment: I.Filippidou (filippidoui@aueb.gr)

Assignment 1

The goal of this assignment is to implement a simple workflow that will assess the similarity between bank customers and suggest for any input customer a list of his/her 10 most similar other customers. In order to calculate the similarity between customers you will first have to compute the dissimilarity for every given attribute as discussed in lecture "Measuring Data Similarity". In order to fulfill this assignment, you will have to perform the following tasks:

1) Import and pre-process the dataset with bank customers

You will download the bank.csv dataset from e-class. This dataset is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls in order to access if the product (bank term deposit) would be (or not) subscribed. The dataset includes 43192 bank customer profiles with 10 attributes each. The class attribute should be ignored. The last attribute is an array containing the bank products (1-20) each customer has. Full description for the dataset and the attributes is provided in the bank-names.txt file. For any numerical missing values, you should replace them with the average value of the attribute (keeping the integer part of the average).

2) Compute data (dis-)similarity

In order to measure the similarity between the bank customers you could form the dissimilarity matrix for all given attributes. As described in lecture "Measuring Data Similarity", for every given attribute you first distinguish its type (categorical, ordinal, numerical or set) and then compute the dissimilarity of its values accordingly. For set

similarity use the Jaccard similarity between sets. Then, you can calculate the average of the computed dissimilarities in order to form the dissimilarity over all attributes. Depending of the machine used to implement this assignment you should decide whether is feasible to compute the dissimilarity matrices or have the computations performed on-the-fly for a pair of customers.

3) Nearest Neighbor (NN) search

Using the dissimilarities computed as discussed in the previous step, you will calculate the 10-NN (most similar) customers for the customers with ids listed below (customer id=line number in the csv file starting from line 2):

1230, 5032, 10001, 24035, 28948, 35099, 37693, 39543, 40002, 42192

For this task your script must take as input the customer-id and return the list of her 10 nearest neighbors (excluding the given customer)

Assignment handout:

1) A report (pdf) describing in detail any processing and conversion you made to the original data and the reasons it was necessary. The report will also contain examples of how to use your script and its output to the list of customers provided at step 3. The first page of the report should clearly state the names and student ids of the members of the group.
2) The program/script you implemented for calculating the dissimilarity matrix. Implementation can be done in any programming language and should be accompanied by the necessary comments and remarks.
3) The pdf and the required programs/scripts should be uploaded to eclass until the assignment deadline. You should create a compressed (e.g. zip/tar) file containing the report, your code and any other files required for executing your script (you do not need to include the original dataset). The name of the compressed file should include the student ids of the members of the group.